Traffic Shift - Avoiding Disasters at Scale

Michael Kehoe
Staff SRE
LinkedIn

Anil Mallapur
SRE
LinkedIn
Overview

LinkedIn Architectural Overview

Fabric Disaster Recovery

Questions
World’s largest professional network

467+ million members

200+ Countries
Who are we?

Production-SRE team at LinkedIn

- Assist in restoring stability to services during site critical issues
- Developing applications to improve MTTD and MTTR
- Provide direction and guidelines for site monitoring
- Build tools for efficient site issue troubleshooting, issue detection & correlation
Terminologies

- **Fabric/Colo**: Data Center with full application stack deployed
- **PoP/Edge (termination)**: Entry point to LinkedIn network (TCP/SSL)
- **Load Test**: Planned stress testing of data centers
2017

4 Data Centers

13 PoPs

1000+ services
What are Disasters?

- Service Degradation
- Infrastructure Issues
- Human Error
- Data Center on Fire
One solution for all disasters

Traffic Shift - Reroute user traffic to different datacenters without any user interruption.
Whaaaaat ?
Request

ATS

EDGE

Request

DC1 in cookie

FABRIC

DC1

DC2

Got DC2 as primary colo for user

If not cookie in header

Gets primary colo for user

Stickyrouting Service
How StickyRouting assigns users to a colo?

Capacity of Fabric

Geographic distance to users

Offline job to assign colo to users
Advantages of sticky routing

- Less latency for users
- Store data where it’s necessary
- Provides precise control over capacity allotment
When to TrafficShift?

Impact Mitigation

Planned Maintenance

Stress Test
Site Traffic and Disaster Recovery

- Traffic stops being served to offline fabrics
- Traffic is shifted to online fabrics

**Distributed Load**
- US-East: 0%
- US-West: 50%
- US-Central: 50%
- APAC: 0%
What is Load Testing?

USW → Target Data Center

- 3 times a week
- Peak hour traffic
- Fixed SLA
Load Testing

US-West

Target

US-East

50%

Traffic Percentage
Benefits of Load Test

- Capacity Planning
- Leverage production traffic to stress test services
- Identify bugs in production
- Confidence in Disaster Recovery
Big Red Button

Kill switch (No Kidding)

Failout of a datacenter and PoP in less than 10 minutes

Minimal user impact
Key Takeaways

- Design infrastructure to facilitate disaster recovery
- Stress test regularly to avoid surprises
- Automate everything to reduce time to mitigate impact
Questions?
Edge
Failout
LinkedIn’s PoP Architecture

- Using IPVS - Each PoP announces a unicast address and a regional anycast address
  - APAC, EU and NAMER anycast regions
- Use GeoDNS to steer users to the ‘best’ PoP
- DNS will either provide users with an anycast or unicast address for www.linkedin.com
  - US and EU members is nearly all anycast
  - APAC is all unicast
LinkedIn’s PoP DR

- Sometimes need to fail out of PoP’s
  - 3rd party provider issues (e.g. transit links going down)
  - Infrastructure maintenance

- Withdraw anycast route announcements

- Fail healthchecks on proxy to drain unicast traffic